ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ
INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT





УДК 910.31 https://doi.org/10.23947/2687-1653-2021-21-4-346-363

Применение инструментов машинного обучения и интеллектуальный анализ данных в отношении баз данных с небольшим количеством записей



Хуберт Аныш ✉
Варшавский технологический университет (г. Варшава, Республика Польша)
✉ h.anysz@il.pw.edu.pl

Использование инструментов интеллектуального анализа данных и машинного обучения становится все более распространенным явлением. Их полезность особенно заметна в случае больших наборов данных, когда информация, которую необходимо найти, или новые взаимосвязи извлекаются из информационного шума. Развитие этих инструментов означает, что исследуются наборы данных с гораздо меньшим количеством записей, обычно связанных с конкретными явлениями. Такая специфика чаще всего приводит к невозможности увеличения количества случаев, а это может облегчить поиск зависимостей в изучаемых явлениях. В статье рассмотрены особенности применения выбранных инструментов к небольшим наборам данных. Предприняты попытки представить методы подготовки данных, методы расчета производительности инструментов с учетом специфики баз данных с небольшим количеством записей. Предложены избранные автором методики, которые помогли выйти из тупика в расчетах, т. е. получить результаты, намного хуже ожидаемых. Необходимость применения методов повышения точности прогнозов и точности классификации была вызвана небольшим количеством анализируемых данных. Эта статья не является обзором популярных методов машинного обучения и интеллектуального анализа данных, тем не менее собранный и представленный материал поможет читателю сократить путь к получению удовлетворительных результатов при применении описанных вычислительных методов.

Ключевые слова: машинное обучение, интеллектуальный анализ данных, искусственные нейронные сети, ассоциативный анализ, автоматическая классификация.

Для цитирования: Аныш, Хуберт. Применение инструментов машинного обучения и интеллектуальный анализ данных в отношении баз данных с небольшим количеством записей / Хуберт Аныш // Advanced Engineering Research. — 2021. — Т. 21, № 4. — С. 346–363. https://doi.org/10.23947/2687-1653-2021-21-4-346-363



Machine Learning and data mining tools applied for databases of low number of records

Hubert Anysz 
Warsaw University of Technology (Warsaw, Poland)
✉ h.anysz@il.pw.edu.pl

The use of data mining and machine learning tools is becoming increasingly common. Their usefulness is mainly noticeable in the case of large datasets, when information to be found or new relationships are extracted from information noise. The development of these tools means that datasets with much fewer records are being explored, usually associated with specific phenomena. This specificity most often causes the impossibility of increasing the number of cases, and that can facilitate the search for dependences in the phenomena under study. The paper discusses the features of applying the selected tools to a small set of data. Attempts have been made to present methods of data preparation, methods for calculating the performance of tools, taking into account the specifics of databases with a

small number of records. The techniques selected by the author are proposed, which helped to break the deadlock in calculations, i.e., to get results much worse than expected. The need to apply methods to improve the accuracy of forecasts and the accuracy of classification was caused by a small amount of analysed data. This paper is not a review of popular methods of machine learning and data mining; nevertheless, the collected and presented material will help the reader to shorten the path to obtaining satisfactory results when using the described computational methods.

Keywords: machine learning, data exploration, artificial neural networks, association analysis, automatic classification

For citation: Hubert Anysz. Machine Learning and data mining tools applied for databases of low number of records. Advanced Engineering Research, 2021, vol. 21, no. 4, pp. 346–363. https://doi.org/10.23947/2687-1653-2021-21-4-346-363

Введение. В эпоху всеобщего доступа в Интернет все больше и больше устройств взаимодействуют друг с другом или с централизованными базами данных. Рекламодатели превосходят друг друга в эффективности персонализированной рекламы. Все это заставляет бурно развиваться группу инструментов, известную как искусственный интеллект. Объем данных, который необходимо обработать, чтобы получить нужную информацию, огромен, поэтому количество публикаций по алгоритмам, позволяющим быстро извлекать информацию из информационного шума, очень велико. Наиболее часто при этом приходится сталкиваться с информационной перегрузкой. Ученые из разных областей знаний знакомы с проблемами, связанными с анализом данных. Нередко сбор данных об изучаемых явлениях требует дорогостоящих устройств, установок и испытаний. Само исследование также может быть длительным. Это означает, что в научно-исследовательских базах данных о причинах и следствиях анализируемых явлений часто может содержаться всего несколько десятков или несколько сотен записей. Преимущества инструментов машинного обучения и интеллектуального анализа данных, в т. ч. возможности поиска значительных зависимостей между многомерными входными и выходными данными, дают возможность исследователям использовать эти инструменты для определения ранее не обнаруженных взаимосвязей изучаемых процессов и явлений. Недостаточное количество записей в созданной базе данных, описывающих какое-либо явление, может снизить ценность полученных результатов анализа. В статье представлены разработки автора, в которых инструменты машинного обучения и интеллектуального анализа данных использовались для исследования материалов и анализа процессов, когда количество входных данных было большим по сравнению с количеством выполненных тестов (т. е. записей в базе данных). Собранные примеры приложений были расширены за счет включения методов подготовки данных и методов оценки точности прогнозов и классификации, чтобы облегчить работу и помочь быстрее достичь ожидаемых результатов людям, которые намерены использовать инструменты машинного обучения для анализа собственных исследований.

Анализ явлений, описываемых многими переменными. Перед любым исследователем непременно встают вопросы: какие входные значения принимать для анализа как влияющие на изучаемое явление, а какие параметры измерять на выходе. Очень полезен бывает статистический подход в исследовании, но он имеет существенный недостаток, который состоит в том, что можно анализировать только одну пару функций. При этом стоит, конечно, определиться, что такое статистика. Согласно [1], «Статистика — это наука о методах проведения статистического обследования и методах анализа его результатов». Предметом же статистического обследования являются отдельный набор объектов, который называется статистическим сообществом (населением), или несколько статистических сообществ. Статистику можно разделить на три основные части: описательная статистика, распределение случайных величин и статистический вывод (рис. 1) [2].



Рис. 1. Статистические управления [2]

В случае, когда на результат процесса влияет множество переменных, с помощью статистических методов очень сложно найти такие комбинации значений входных переменных, которые существенно влияют на изменчивость выходных данных. И самое востребованное — как эффективно управлять процессом или явлением, чтобы на выходе получить желаемый результат. С помощью инструментов интеллектуального анализа данных гораздо проще найти взаимосвязи между многомерными входными и выходными данными. Интеллектуальный анализ данных очень точно определяется самим названием книги [3], которое можно перевести как «Обнаружение знаний из данных». Существует определение интеллектуального анализа данных, сформулированное в 2001 году как анализ (часто огромных) наборов данных наблюдений с целью обнаружения неожиданных взаимосвязей и обобщения данных оригинальным способом, чтобы они были понятны и полезны своему владельцу [4–5]. Для этих нужд разрабатываются методы и алгоритмы, благодаря которым поиск вышеупомянутых соединений происходит быстрее и эффективнее. Методы интеллектуального анализа данных можно разделить на:

— обнаружение ассоциаций (правила ассоциации);
— классификация и прогнозирование;
— группировка;
— анализ последовательности и времени;
— обнаружение характеристик;
— интеллектуальный анализ текстовых и полуструктурных данных;
— изучение контента, размещенного в Интернете;
— изучение графиков и социальных сетей;
— интеллектуальный анализ мультимедийных и пространственных данных;
— обнаружение особенности [6].

На этой основе были разработаны методы, обычно называемые искусственным интеллектом, благодаря которым выполняются наиболее часто выбираемые задачи интеллектуального анализа данных. Несмотря на развитие информационных технологий и возрастающую вычислительную мощность компьютеров, до сих пор практически невозможно проверить все возможные комбинации многомерного ввода и вывода сложной системы [7]. Использование методов искусственного интеллекта тем более оправдано, чем сложнее проблема и не известны механизмы, управляющие ею, как показано на рис. 2.



Рис. 2. Предлагаемые условия использования методов искусственного интеллекта [7]

Существует множество методов и приемов искусственного интеллекта (включая искусственные нейронные сети, метод K-ближайших соседей, случайный лес, деревья решений), и они все еще разрабатываются[1]. Популярность, которую можно прочесть как полезность приложений одного из инструментов искусственного интеллекта — искусственных нейронных сетей очень заметна, например, на

[1] StatSoft. Internetowy Podręcznik Statystyki. URL: https://www.statsoft.pl/textbook/stathome.html (accessed: 10.07.2020).

основе [8]. Вместо строгого поиска возможных комбинаций используется метаэвристика. Если необходимо использовать вышеупомянутые инструменты, следует решить, пользоваться ли специализированным программным обеспечением или создать его самостоятельно с помощью общедоступных модулей (так называемых «движков»), реализующих алгоритмы искусственного интеллекта. Независимо от принятого решения в основе будут лежать данные, которые будут анализироваться.

Подготовка данных. Можно выделить следующие этапы подготовки данных к анализу:

— очистка данных;

— интеграция данных;

— выбор данных;

— консолидация и преобразование данных [6].

Такая подготовка должна выполняться независимо от размера базы данных. Для небольших наборов данных их правильная подготовка даже более важна, чем для больших. Примером может служить сравнение двух наборов данных: один с 10 000 записей, а другой со 100 записями, где 5 % записей относятся к повторяющемуся явлению (повторяемость еще не обнаружена). Когда две записи содержат ошибочные данные, то в первом случае можно найти повторяемость в 4,8 % случаев вместо 5,0 %. Во втором случае повторяемость обнаруживается только в 3,0 %. Разница существенная.

Очистка и интеграция данных. При очистке данных из базы удаляются в основном записи, содержащие неполные данные. В больших базах данных удаление, например, двух записей существенно не повлияет на результаты, полученные на последующих этапах. При небольшом количестве наборов данных потеря даже одной записи может существенно повлиять на полученные результаты анализа. По этой причине отсутствующие значения не могут быть заменены, например, средним для всей генеральной совокупности (один из методов увеличения данных) или ее части (аналогичной описанию в записи, которую следует удалить), как это делается для больших баз данных Причина та же, что и описана выше — замена одного отсутствующего признака в описании явления может существенно изменить результаты, если анализируется небольшой набор данных. Однако записи, удаленные в процессе, не должны удаляться безвозвратно. На последующих этапах может оказаться, что в окончательно принятой модели данная особенность не будет учтена, а изначально удаленная запись будет содержать полные данные — это будет полезно для анализа.

Второй важный этап очистки данных — это статистический анализ каждой характеристики (столбцов в базе данных) отдельно и ее корреляция с выходными данными. Рекомендуется представлять статистику основных характеристик анализируемого процесса (количество записей, среднее арифметическое, медиана, минимальное и максимальное значение, стандартное отклонение, квартили значений характеристик) также и для функции или функций, описывающих выходные данные. Диаграммы «рамка-усы» очень удобочитаемы (рис. 3).



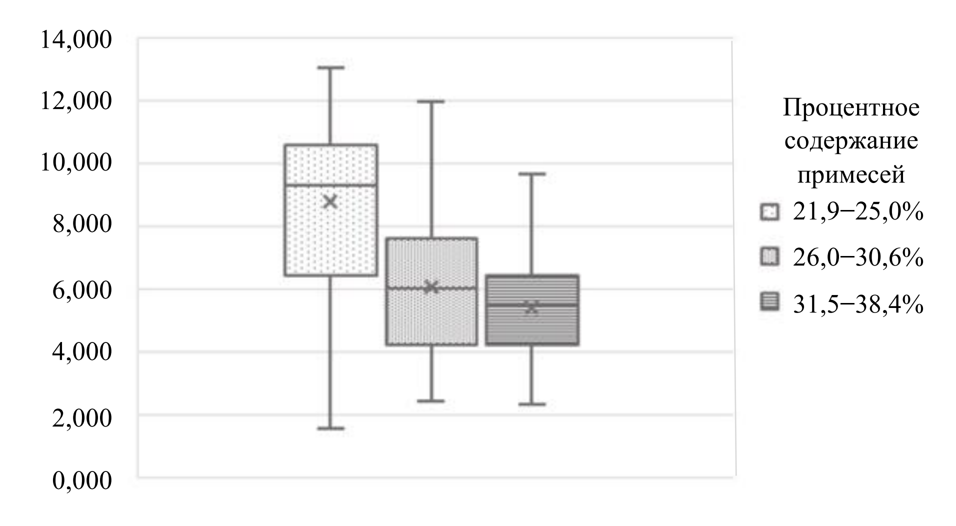


Рис. 3. Пример диаграммы типа «кадр-усы» [9]

На таком графике легко прочитать, например, что для 50 % образцов с 21,9–25,0 % содержания глины с пылью прочность была выше 9 МПа, но при этом минимальная прочность для этого типа образцов была ниже 2 МПа. Для таких образцов прочность была ниже 6,5 МПа. Анализ базовой статистики может облегчить решение об исключении из анализа записей (т. е. образцов или исследуемых явлений), для которых измеренные значения несовместимы со всеми другими случаями. Значительное несоответствие может быть результатом ошибочного измерения или того факта, что на измерение повлиял другой фактор, который вообще не принимался во внимание (он не учитывался и не измерялся ни в одном из случаев). По этим причинам все записи, отклоненные из базы данных, должны быть описаны, а также должны быть указаны причины отклонения [10].

Другой случай. Если обнаружено, к примеру, что решение об отклонении записи может быть принято только после того, как будут выполнены все или часть вычислений. Об этом идет речь в статье [11]: на основе наборов из 95 ускорений (a) стандартизированного молота (ударяющего по испытуемому стальному элементу), измеряемых каждые 0,01 мс с использованием искусственных нейронных сетей, была предпринята попытка отнести испытываемый стальной элемент к одному из девяти классов (рис. 4, 5).

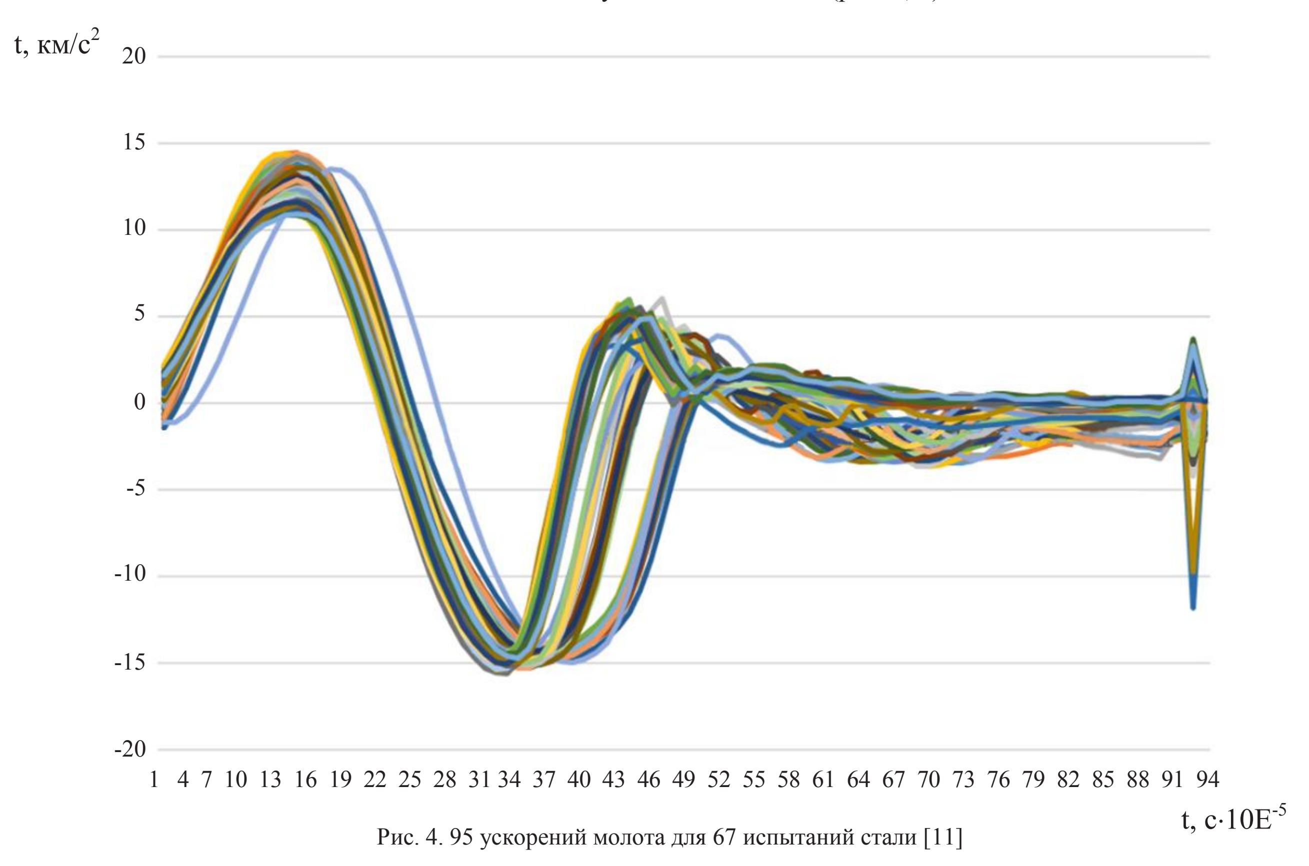


Рис. 4. 95 ускорений молота для 67 испытаний стали [11]

Анализируя ускорения на рис. 4, можно сказать, что один из тестов во временном диапазоне от 0,01 до 0,31 мс отстает от других, но по-прежнему ведет себя как другие образцы. Только предварительная классификация по четырем группам марок стали показала, что через 0,31 мс в испытании № 29 были получены результаты, которые показывают резко отклоняющийся характер результатов также через 0,31 мс (рис. 5). В испытании № 29 был исследован стальной образец, для которого во всех других испытаниях ускорение изменило знак с отрицательного на положительный между 0,424 и 0,450 мс. Для испытания № 29 знак ускорения изменился в пределах 0,460–0,495 мс, то есть за время, подходящее для другой группы марок стали. Только этот вывод позволил достаточно хорошо обосновать отклонение из анализов испытания № 29. В результате была повышена первоначально полученная точность классификации до девяти марок стали по результатам 67 испытаний, равная 80 %, до 95 % (после отказа от теста № 29 и повторного использования искусственных нейронных сетей).

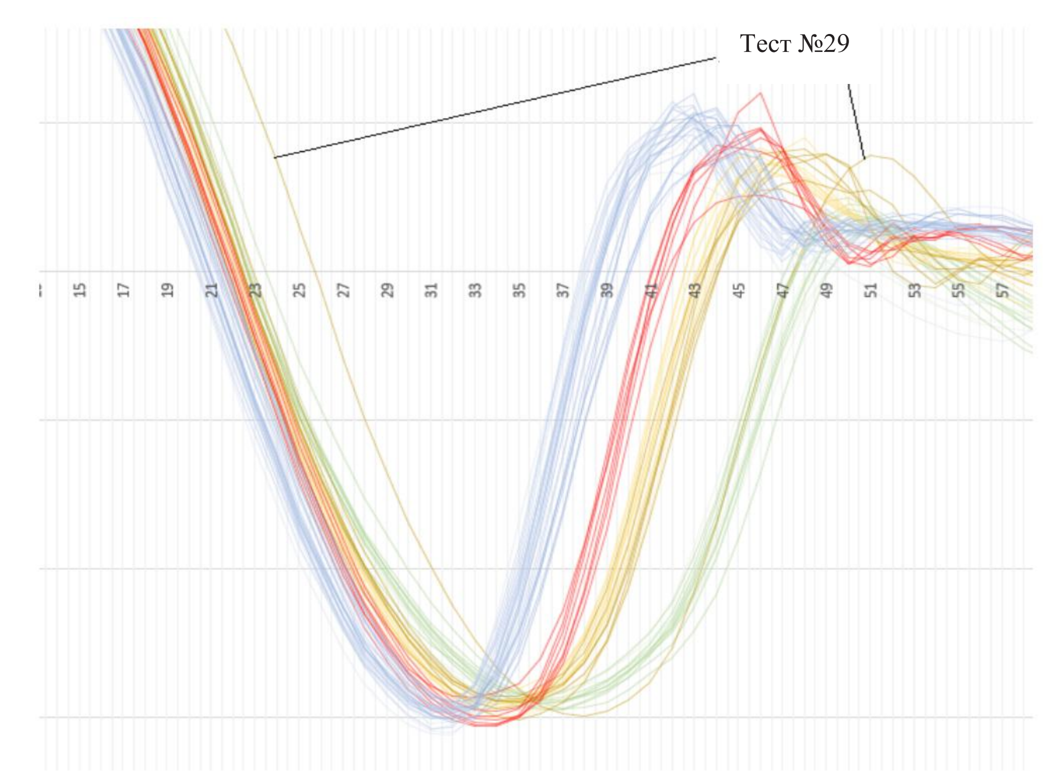


Рис. 5. Фрагмент диаграммы с рис. 4 с предварительной классификацией на четыре группы классов сталей и выпадающим тестом № 29 также через 0,31 мс [11]

Интеграция данных — это объединение данных об одном и том же явлении из разных источников в одну базу данных. Пример интеграции содержится в работе [12], в которой прогнозируется задержка строительства участков скоростных и автомагистралей в Польше. Независимыми переменными, на основании которых делались прогнозы, являются данные о построенных объектах (предусмотренных законом о доступе к информации главного управления национальных дорог и автомагистралей), данные о предприятиях, реализующих эти объекты (собранные в регистрационный суд), Интернет, агентство бизнес-аналитики, макроэкономические данные (с источником в публикациях центрального статистического управления). Значение зависимой переменной (необходимое для «обучения» искусственной нейронной сети) — количество дней, на которое откладывается завершение каждой из проанализированных дорожных инвестиций, искали в публикациях в прессе и в Интернете. Собранная информация была использована для интеграции в базу данных о реализации 128 строительных проектов. В Польше в 2009–2013 годах было построено 156 участков скоростных и автомагистралей, но получить полную информацию о них было практически невозможно. После анализа хода этого строительства были отклонены и те случаи, когда возникали неожиданные нарушения (например, в виде протестов экологов, которые не учитывались в анализах как независимая переменная). Это уменьшило количество дел на 28, но обеспечило полноту, целостность базы данных — основы расчета.

Выбор данных. В больших наборах данных их размер является существенной проблемой — большое количество записей приводит к неэффективной и длительной работе программного обеспечения. В базах данных с небольшими размерами записей программному обеспечению поиска отношений ввода-вывода может быть недостаточно, чтобы найти эти отношения. Бывает, что изучаемое явление можно описать многими параметрами, но в базе данных мало случаев (записей) с описанными параметрами явления. Таким образом, выбор данных означает необходимость выбора лишь нескольких независимых переменных, на основе которых будут выполняться классификация или прогнозирование выходного значения с использованием искусственного интеллекта (также известного как машинное обучение). При выборе независимых переменных может оказаться полезным следующее:

— изучение взаимной корреляции линейных независимых переменных, а также корреляции со значениями на выходе;

— анализ основных компонентов;

— эмпирический поиск оптимального набора независимых переменных.

Корреляционное исследование. Исследование линейной корреляции Пирсона между парами независимых переменных и между каждой из них и зависимой переменной может быть представлено в форме таблицы с числами, а также графически, в виде так называемых «тепловых карт» (пары независимых

переменных) [13]. Переменные наиболее сильно коррелируют положительно, а интенсивный синий цвет на рис. 6 демонстрирует наименьшее значение коэффициента Пирсона. Сильная положительная или отрицательная корреляция, считанная с тепловой карты, не обязывает удалять переменную, сильно коррелированную с другой, это всего лишь предположение, потому что она сильно положительно коррелирует с zn2 (коэффициент корреляции между ними составляет 0,88), и в то же время zn5 не коррелирует с выходом (обозначено как wy, коэффициент корреляции равен 0,03).

	zn1	zn2	zn3	zn4	zn5	wy
zn1	1,00	0,11	-0,58	0,17	0,77	0,90
zn2	0,11	1,00	0,03	-0,30	0,88	0,35
zn3	-0,58	0,03	1,00	0,30	0,52	-0,34
zn4	0,17	-0,90	0,80	1,00	0,41	0,70
zn5	0,77	0,88	0,52	0,41	1,00	0,03
wy	0,90	0,35	0,34	0,70	0,03	1,00

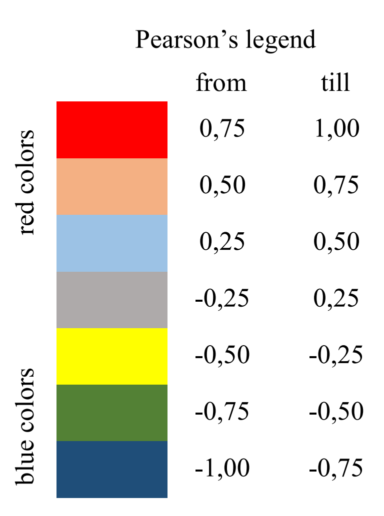


Рис. 6. Примерная «тепловая карта» независимых переменных от n1 до n5 и выходных данных

Хотя вычисляется линейная корреляция, и фактическая связь между независимыми переменными (или независимой переменной и переменной, зависящей от выпуска) может быть не линейной, вычисление этих линейных корреляций часто подсказывает, какие переменные не включать (если есть необходимость их уменьшения). Такая проверка была сделана среди прочего в работах [9, 12]. В [12] количество зависимых переменных было сокращено, а в [9] для анализа была принята новая переменная как сумма значений двух сильно положительно коррелированных независимых переменных (это также было технически оправданно). Переменная, которая имеет сильную отрицательную корреляцию с другой независимой переменной, также может быть удалена из базы данных.

Анализ главных компонентов. Анализ главных компонент (PCA) выполняется для независимых переменных — выходное значение не учитывается[2] В результате получаем рейтинг, показывающий, какая из независимых переменных больше всего влияет на изменчивость наборов независимых переменных. Можно представить, что каждая из независимых переменных осуществляет измерение многомерного пространства. Независимые переменные связаны между собой, они образуют множества (записи в базе данных, описывающие явление). Результатом PCA является ответ на вопрос, какая из независимых переменных наиболее ответственна за то, что расстояния (в многомерном пространстве) между точками (наборы независимых переменных, описанные в записях — координаты точек) являются наибольшими. Переменные, оказывающие наименьшее влияние на разброс данных, это те, которые могут быть удалены из анализа в попытке уменьшить количество независимых переменных. Примеры эффективного применения анализа главных компонентов для повышения производительности инструментов машинного обучения можно легко найти, например, в работах [14–16]. Однако стоит иметь в виду, что PCA не принимает во внимание значение зависимой переменной. Следовательно, нет уверенности в том, что именно независимая переменная, которая также оказывает наибольшее влияние на прогнозируемое значение (зависимая переменная), вызывает наибольшую изменчивость в наборах данных.

Эмпирические исследования. И корреляционное исследование, и анализ главных компонентов не дают абсолютной уверенности в том, был ли выбор независимых переменных оптимальным. Оптимальный — это значит наиболее точный прогноз или максимально возможная доля точных классификаций для базы данных и выбранного инструмента машинного обучения. Инструменты искусственного интеллекта чаще всего применяются, когда их пользователь подозревает, что существует связь между вводом и выводом (между наборами независимых переменных и эффектом их совместного появления — зависимой переменной). Когда эти зависимости невозможно описать строго (функцией многих переменных), когда изучаемые процессы и

[2] StatSoft. Internetowy Podręcznik Statystyki. URL: https://www.statsoft.pl/textbook/stathome.html (accessed 10.07.2020).

явления сложны (рис. 2), то использование инструментов машинного обучения может оказаться единственным способом, чтобы узнать об этом. Поэтому трудно ожидать, что какой-либо вспомогательный инструмент точно укажет, какие из зависимых переменных следует использовать для «обучения» искусственного интеллекта. Следовательно, одним из методов поиска оптимального набора входных данных (зависимых переменных) является эмпирическая проверка результатов инструмента искусственного интеллекта на различных наборах зависимых переменных. Есть два основных способа действия: вперед и назад. Вперед — это означает выбор двух зависимых переменных, на основе которых результаты прогноза или классификации являются наилучшими. Выбрать первый порой очень просто, сложно представить прогноз задержек без указания планируемой продолжительности [12, 17]. Чтобы выбрать вторую переменную, проверяем работу инструмента на каждой созданной паре зависимых переменных (их иногда называют предикторами) [18]. Когда выбрана лучшая пара предикторов, один из оставшихся предикторов добавляется последовательно. Это делается до тех пор, пока добавление какой-либо еще не использованной независимой переменной не улучшит результаты. В обратной процедуре первым шагом является использование всех независимых переменных, а затем последовательное удаление только одной, проверка того, какой предиктор был удален, точность прогноза и классификации увеличилась больше всего. Процедура продолжается до тех пор, пока удаление любого из предикторов не приведет к улучшению результатов.

Консолидация и преобразование данных. Консолидация и преобразование данных состоит в том, чтобы они могли использоваться выбранным инструментом интеллектуального анализа данных [6]. Наиболее распространенной формой преобразования данных является их стандартизация, то есть такое преобразование значений независимых переменных и зависимой переменной, при котором они принимают значения из одного и того же диапазона. Стандартизация данных является результатом необходимости предоставить каждой из независимых переменных «равные условия для игры», которые будут включены в модель машинного обучения. В [12] дана формула:

$$dla\ 1 \le i \le\ k \qquad a_{1i} = \frac{a_{0i}}{max_k(a_{0i})} \tag{1}$$

где k — количество записей в базе данных;

a_{0i} — i-й элемент переменной a до стандартизации;

a_{1i} — i-й элемент переменной a после стандартизации.

Второй широко используемый тип стандартизации данных — это так называемая стандартизация «к нулевому среднему значению и стандартному отклонению единицы», определяемая следующей формулой:

$$a_{1i} = \frac{a_{0i}-\overline{a_0}}{\sigma_a} \tag{2}$$

где $\overline{a_0}$ — среднее арифметическое значение переменной a до стандартизации;

a_{0i} — i-й элемент переменной a до стандартизации;

a_{1i} — i-й элемент переменной a после стандартизации;

σ_{0a} — стандартное отклонение переменной a до стандартизации [7].

Другие типы стандартизации, также нелинейные, можно найти, например, в [19]. Однако следует помнить, что тип стандартизации данных может изменить результаты, полученные с помощью машинного обучения [20]. Таким образом, тип стандартизации данных может быть одним из параметров, которые инструмент настраивает для получения наилучших результатов.

Бинаризация может быть вторым процессом преобразования данных. Она означает преобразование числовых значений переменной только в два значения (например, 0 и 1) по следующей формуле [22]:

$$a_{1i} = \begin{cases} a_{0i} \le p \to 0 \\ a_{0i} > p \to 1 \end{cases} \tag{3}$$

где a_{0i} — i-й элемент переменной a до бинаризации,

a_{1i} — i-й элемент переменной a после бинаризации,

p — параметр, выбранный пользователем.

Бинаризация данных особенно полезна при поиске правил с использованием анализа корзины (также известного как анализ ассоциации)[3]. Этот тип анализа был создан для исследования содержимого корзин клиентов с целью увеличения продаж. Компьютерные программы с модулем анализа корзины работают наиболее эффективно, если переменные являются двоичными (данный товар присутствовал в корзинах покупателя или нет). Можно сформулировать множество научных задач, сама суть которых является двоичной, но в большинстве случаев описание явления включает числа, которые преобразуются в двоичную форму для применения анализа корзины [21–22]. Такое преобразование также может быть выполнено для переменной,

[3] StatSoft. Internetowy Podręcznik Statystyki. URL: https://www.statsoft.pl/textbook/stathome.html (accessed 20.07.2020).

которая может принадлежать нескольким непересекающимся подмножествам. Затем два дихотомических подмножества (являющиеся суммами исходных подмножеств) создаются из первичных подмножеств. Тогда, если a_{0i} принадлежит одному из них, то a_{1i} равно 0, если другой равен 1.

Измерения ошибок. До сих пор автором использовались общие термины, такие как «точность прогнозов», «правильность классификации», которые можно назвать качеством или эффективностью инструментов искусственного интеллекта. Однако, если проанализировать способы повышения качества их работы, то необходимо определить ошибки в результатах действия инструментов машинного обучения и итогах интеллектуального анализа данных.

Ошибки прогноза. Инструменты машинного обучения в основном служат двум целям: для прогнозирования значений (регрессия) и для автоматической классификации. Использование одних и тех же общепринятых мер ошибок облегчает понимание работы, но также позволяет легче оценить ценность прогноза. Предполагается, что анализируется абсолютное значение ошибки. Следовательно, абсолютная ошибка (AE) [7, 18] может быть определена как

$$AE = \left|\hat{b} - b\right| \tag{4}$$

где $\hat{b}$ — прогнозируемое значение;

b — фактически наблюдаемое значение.

Относительная ошибка, выраженная как абсолютная ошибка в процентах (APE), определяется как

$$APE = \left|\frac{\hat{b}-b}{b}\right| * 100\% \tag{5}$$

Чтобы иметь возможность оценить качество прогнозов, сделанных с помощью инструмента искусственного интеллекта (например, искусственной нейронной сети), некоторые данные не используются в процессе «обучения». После построения модели вводится этот набор данных, называемый проверочной выборкой, и машина делает прогнозы. Таким образом, прогнозируемые значения составляют дюжину, несколько дюжин или больше. Затем для оценки качества прогнозов можно рассчитать среднюю абсолютную процентную ошибку (MAPE):

$$MAPE = \frac{\sum_{i=1}^{n}\left(\left|\frac{\widehat{b_i}-b_i}{b_i}\right| * 100\%\right)}{n} \tag{6}$$

где *n* — размер валидационной выборки.

Наиболее распространенной мерой погрешности (проверочный тест) является среднеквадратичная ошибка (MSE), определяемая как

$$MSE = \frac{\sum_{i=1}^{n}\left(\widehat{b_i}-b_i\right)^2}{n} \tag{7}$$

При решении задач регрессии большинство инструментов машинного обучения благодаря использованию эвристических алгоритмов ищет отображение ввода и вывода, которое минимизирует MSE. При указании качества полученных прогнозов наиболее распространенными являются MSE или MAPE (или оба). Следует отметить, что в случае MAPE не имеет значения, рассчитывается эта ошибка для стандартизованных или реальных значений — MAPE то же самое. В случае с MSE дело обстоит иначе. Эта ошибка чаще всего имеет разные значения для стандартизированных прогнозов и для прогнозов, преобразованных в истинные значения (без стандартизации). Следовательно, необходимо указать, для каких значений MSE были рассчитаны. Сравнение точности прогнозов (различных процессов, явлений с разными инструментами) на основе MSE оправдано, если MSE рассчитывается для стандартизованных значений. С другой стороны, с точки зрения практического применения важнее MAPE или максимальное значение AE. Полезность получаемых прогнозов также является важным вопросом [23]. Прогноз косвенных затрат на строительство со средней относительной погрешностью 6% можно считать очень точным и полезным прогнозом, но те же 6 % MAPE для прогнозов фондовой биржи делают их бесполезными [24]. Следовательно, размер ошибки, полученной в прогнозах, также следует оценивать с точки зрения полезности для лиц, принимающих решения, использующих прогнозы.

Меры точности классификации. При использовании инструментов машинного обучения для автоматической классификации отнесение случая к неправильному классу можно оценить двумя способами. Во-первых, просто как ошибка. Однако одной информации о том, что инструмент правильно классифицирует 90 % случаев, может оказаться недостаточно. Если есть несколько классов, которым назначены отдельные случаи (описанные в базе данных) (например, восемь), может случиться так, что для пяти классов классификация будет на 100 % правильной, а 10 % ошибок относятся к другим трем классам. Поэтому качество результатов классификации оценивается так называемой матрицей ошибок, пример которой представлен в таблице 1.

Таблица 1

Результаты классификации тендерных процедур по валидационной выборке [25]

	Класс: свободный от сговора	**Класс: подозрение в сговоре**	**Класс: сговор очень вероятный**	**Всего: для всех классов**
Численность в валидационной выборке	52	14	4	70
Количество правильных классификаций	50	10	3	63
Количество неправильных классификаций	2	4	1	7
Доля правильных классификаций, %	96,15	71,43	75,00	90,00
Доля неправильных классификаций, %	3,85	28,57	25,00	10,00

Значительные различия в точности классификации отдельных подмножеств могут способствовать дальнейшему поиску еще более точной модели классификации. Матрица ошибок также может содержать информацию, к какому неправильному классу была неправильно отнесена данная запись из выборки проверки. Это влияет на вывод, основанный на прогнозах. Проанализируем пример из таблицы 1 со следующими предположениями:

— два неправильно классифицированных делопроизводства из класса «свободный от сговора», отнесенных автоматическим классификатором к классу «подозрение в сговоре»;

— четыре неправильно классифицированных случая из класса «подозрение в сговоре» были отнесены автоматическим классификатором к классу «весьма вероятный сговор»;

— один неправильно классифицированный случай из класса «очень вероятный» был отнесен автоматическим классификатором к классу «подозрения в сговоре».

В результате анализа можно констатировать, что до тех пор, пока классификатор не отнесет данную процедуру к классу «свободный от сговора», можно быть уверенным, что это производство не связано со сговором. Все дела, отнесенные к этому классу, были правильно классифицированы автоматическим классификатором (несмотря на точность классификации менее 100 %). Этот эффект использовался, например, в [11]. Следовательно, стоит проанализировать, к каким классам автоматически был отнесен данный случай.

Ошибкой классификации в медицинских приложениях является разделение ошибок только на два класса, где так же важно не вводить лекарства здоровому человеку, как и не отказывать в лечении действительно больному человеку (принимая его за здорового) (рис. 7).

		Класс присваивается классификатором	
		Положительный	Отрицательный
Отсортированный класс	Положительный	Количество положительных верных, определенных, как TP	Количество отрицательных неверных, определенных, как FN
	Отрицательный	Количество положительных неверных, определенных, как FN	Количество отрицательных верных, определенных, как TN

Рис. 7. Матрица ошибок классификации на два класса; серый фон указывает на правильную классификацию[4] [26]

[4] PQStat Statystyczne Oprogramowanie Obliczeniowe. URL: https://pqstat.pl/?mod_f=diagnoza (accessed 10.07.2020).

Для n классифицированных случаев выполняется следующее равенство:

$$n = TP + FP + TN + FN \quad (8)$$

Для интерпретации матрицы ошибок в форме, представленной на рис. 7, используются понятия точности, прецизионности, чувствительности, специфичности, определяемые следующими уравнениями[5] [26]:

$$\text{точность} = \frac{TP+TN}{n} \quad (9)$$

$$\text{прецизионность} = \frac{TP}{TP+FP} \quad (10)$$

$$\text{чувствительность} = \frac{TP}{TP+FN} \quad (11)$$

$$\text{специфичность} = \frac{TN}{FP+TN} \quad (12)$$

Однако следует помнить, что указанные выше индикаторы применимы лишь в том случае, если инструмент классифицирует только два класса.

Определение значимости обнаруженных ассоциативных правил. Благодаря использованию анализа корзины — одного из инструментов интеллектуального анализа — в данных обнаруживаются правила, которые можно записать как

$$b \to h \quad (13)$$

где b — предшественник; h — преемник правила.

Такое правило читается так: если был предшественник, то был и преемник. И предшественник, и последующий могут состоять из нескольких переменных, но наиболее часто ищутся правила, в которых предшественник описывается многими переменными, а последующий — одним (например, если давление упало утром, а температура в полдень превысила 30℃, потом днем была гроза). Такое правило (как в примере выше) не всегда работает. Следовательно, мерами качества обнаруженных правил являются не меры ошибок, а три параметра (пропорции) [6, 9], благодаря которым можно легко определить, что правило будет проверено в случае нарушения предшественника [3, 4, 6]:

— поддержка — помечено как *sup* (от англ. support);

— уверенность — обозначается как *conf* (от англ. confidence);

— увеличение — отмеченный как *lift* (от англ. lift).

Поддержка определяется следующим образом:

$$sup(b \to h) = \frac{n(b \to h)}{N} \quad (14)$$

где $n(b \to h)$ — количество случаев, в которых возникновение предшественника сопровождалось появлением наследника; N — количество всех случаев в базе данных.

С другой стороны, достоверность правила определяется следующим образом:

$$conf(b \to h) = \frac{n(b \to h)}{n(b)} \quad (15)$$

где $n(b)$ — количество случаев, в которых было отмечено появление предшественника.

Не менее важно приращение правила. Если его значение меньше 1, это означает, что найденное правило не объясняет появления наследника. Приращение определяется следующим образом:

$$lift(b \to h) = \frac{conf(b \to h)}{P(h)} \quad (16)$$

где $P(h)$ — вероятность преемника (независимо от того, был ли предшественник или нет).

Лучшему пониманию оценки качества обнаруженных ассоциативных правил будет способствовать пример, в котором в 12 процессах было сделано 10 наблюдений за возникновением предшественника и последующего (рис. 8). Для правила «если предшественник, то наследник» были рассчитаны поддержка, достоверность и приращение (для каждого из процессов), которые представлены в таблице 2.

[5] PQStat Statystyczne Oprogramowanie Obliczeniowe, там же.

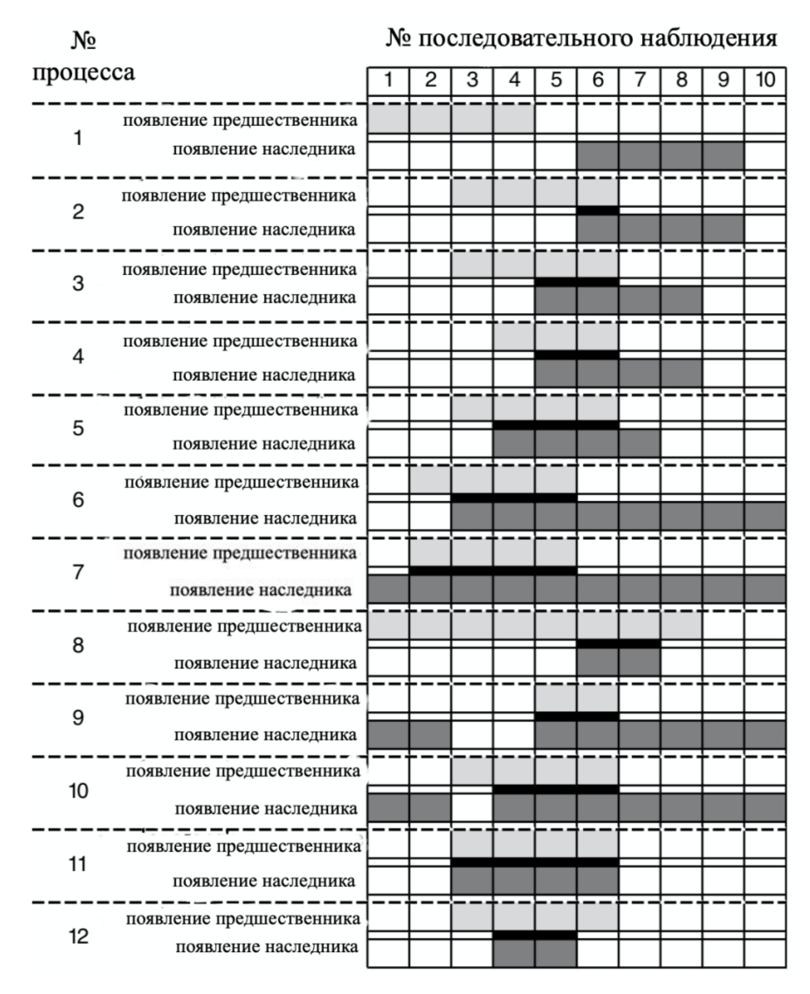


Рис. 8. Наблюдения за появлением предшественника и преемника в 12 процессах

Сравнивая процессы 5 и 6, следует отметить, что высокая степень достоверности правил не всегда важна. В процессе 6 преемник почти всегда присутствует, и обнаруженное правило не объясняет возникновение преемника (подъем<1). Это не относится к процессу 8. Приращение указывает на важность правила, в то время как его достоверность низкая. Тем не менее, каждое наблюдение за преемником сопровождается наблюдением за предшественником. В случае процесса 8, следовательно, стоит уточнить предшественник (например, добавив еще одну переменную). В этом случае, вероятно, другой параметр (пока не включенный в предшественник) влияет на наличие преемника (в процессе 8).

Таблица 2

Оценка правил с использованием *sup, conf* и *lift* для процессов на рис. 8

№ процесса	**Поддержка (Sup)**	**Уверенность ($Conf$)**	**Увеличение ($Lift$)**
1	0,00	0,00	0,00
2	0,10	0,25	0,63
3	0,20	0,50	1,25
4	0,20	0,67	1,67
5	0,30	0,75	1,88
6	0,30	0,75	0,94
7	0,40	1,00	1,00
8	0,20	0,25	1,25
9	0,20	1,00	1,25
10	0,30	0,75	0,83
11	0,40	1,00	2,50
12	0,20	0,50	2,50

Важные аспекты применения выбранных инструментов искусственного интеллекта и интеллектуального анализа данных к небольшим базам данных

Количество записей в базе данных и сложность инструмента. В статистике чаще всего предполагается, что небольшой размер выборки составляет менее 30 случаев, однако можно обнаружить, что предельное число, за пределами которого мы не можем говорить о малой выборке, составляет 100 [27–28]. Инструментам искусственного интеллекта нужны полные наборы данных (входы и выходы), чтобы иметь возможность найти наиболее точный способ преобразования одного в другой. Чем сложнее проблема, тем больше требуется наборов данных (записей в базе данных). В глубоком обучении, когда инструмент преподается не с наборами чисел, а с файлами (графикой, аудио, текстом), необходимы тысячи наборов данных. В [29] более 4000 стандартизированных изображений использовалось для прогнозирования прочности на сжатие. В других исследованиях образцов было немного (например, в [11] всего 66). Есть указание на то, что для искусственных нейронных сетей количество связей между нейронами должно быть в 10 раз меньше, чем количество записей в базе данных [30]. Небольшие базы данных требуют большего количества вычислительных испытаний и более точной настройки используемых инструментов. Тем не менее, правило, что чем сложнее проблема и ее модель, тем больше наборов данных требуется для обучения инструмента, остается в силе. Использование сложных моделей (например, искусственных нейронных сетей с более чем одним скрытым слоем и множеством нейронов в скрытых слоях) в небольших базах данных чаще всего приводит к ошибкам (прогнозирования или классификации), намного большим, чем в моделях с меньшей сложностью самого инструмента. Отсюда популярность методов уменьшения количества независимых переменных, описанных в разделе 3. Когда количество наборов данных слишком мало, уменьшение количества независимых переменных чаще всего повышает точность прогнозов и классификации.

Тип вывода и качество прогноза и классификации. В [30] можно найти предположение, что искусственные нейронные сети могут более точно определить, будет ли, например, предсказанное значение больше, чем значение, данное пользователем сети. Ссылаясь на требование полезности прогноза [23], если полученные прогнозы недостаточно точны (т. е. ошибки прогноза слишком велики), можно решить, требуется ли точное значение. Например, при прогнозировании прочности материала возможно вместо прогноза лишь сообщить, что прочность не будет ниже предполагаемой прочности, указанной пользователем. То же самое и с проблемой классификации. Автоматическая классификация стали на 9 марок на основе всего 66 записей в базе данных не позволила получить точность классификации выше 80% [11]. Затем один процесс классификации был заменен восьмым, в результате чего результаты испытаний стали были разделены на две дихотомические подгруппы, как показано на рис. 9.

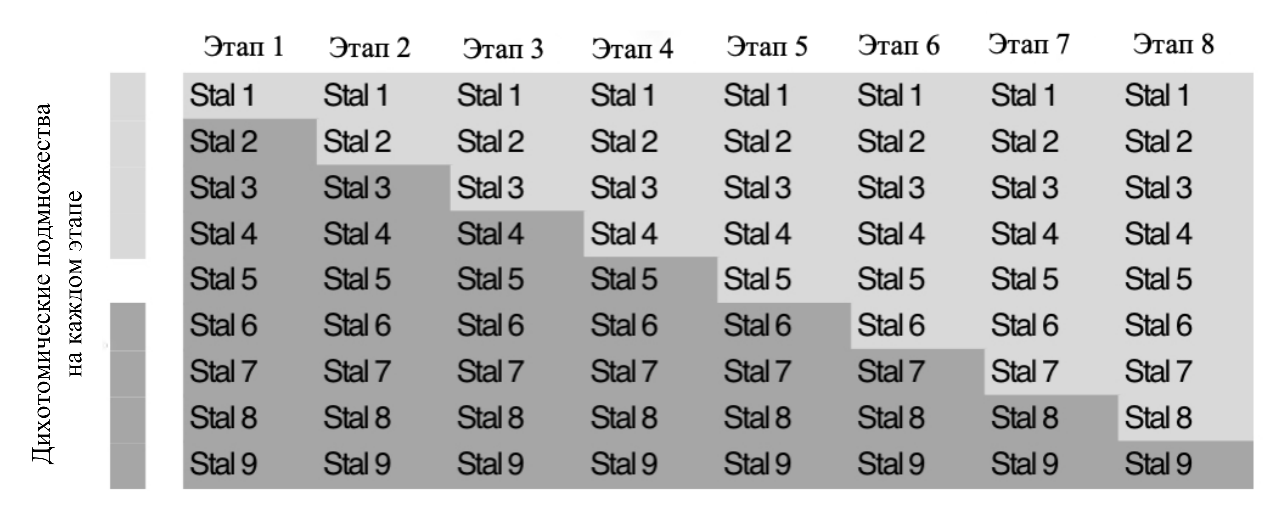


Рис. 9. Восьмиэтапный процесс классификации на два дихотомических подмножества (на каждом этапе классификации) [11]

Автоматическая классификация только двух подмножеств была проведена с помощью более простой модели (со сложностью, соответствующей количеству испытаний стали, т. е. 66). Использованный восемь раз процесс позволил повысить точность классификации с 80 до 95 %. Из первоначально выбранных 12 предикторов осталось только 6 [12]. Это позволило сделать прогноз с достаточно малой ошибкой, чтобы

модель была полезной. Компьютерные программы позволяют прогнозировать, например, две зависимые переменные одновременно, но по причинам, описанным выше, лучшие результаты (меньшие ошибки) получаются при прогнозировании двух зависимых переменных отдельно, с помощью двух моделей, хотя результаты получаются на одном и том же наборе данных.

Модификация вывода. Рассмотрим случай, когда на начальном этапе расчетов не получаются достаточно точные прогнозы или классификации. В дополнение к действиям, связанным с данными и сложностью в инструменте машинного обучения, использованном выше, можно рассмотреть возможность поиска другой зависимой переменной, той, на основе которой можно будет вычислить строго зависимую переменную, которую необходимо найти. При небольшом количестве записей в базе данных это может упростить модель, что, в свою очередь, может повысить точность прогнозов. Этот эффект был использован в [31], где прогнозирование с помощью искусственной нейронной сети является лишь частью процесса проектирования состава смеси. Еще одна процедура, которая может уменьшить ошибки прогнозов, — это прогноз относительного значения (вместо абсолютного значения). При оптимизации работы модели в [12] прогнозируемая задержка, выраженная в днях, была заменена задержкой, выраженной как пропорция количества дней задержки к запланированному количеству дней строительства. В указанном случае это не уменьшило ошибок в прогнозах. Другая возможность изменить тип вывода (т. е. прогнозируемую зависимую переменную) — заменить одно число несколькими значениями функций принадлежности, рассчитанных на основе теории нечетких множеств [32]. В [17] вместо количества дней задержки построения на выходе из искусственной нейронной сети использовались три значения функции принадлежности множествам: низкая задержка, средняя задержка и большая задержка. После уточнения прогнозируемых значений оказалось, что ошибки прогноза были меньше, чем при прогнозировании количества дней задержки [12]. То же самое было предпринято в [33] путем прогнозирования значений функции принадлежности на первом этапе вычислений, и только на втором этапе на основе этих прогнозов случаи были разделены на три подмножества. Однако в этом случае прямое использование искусственной нейронной сети в качестве классификатора привело к повышению точности классификации.

Гибридные инструменты. При небольшом количестве кейсов в базе данных используемый инструмент машинного обучения не может быть очень сложным, поскольку слишком мало кейсов для успешного обучения модели. Гибридные модели могут быть средством от слишком большой ошибки прогноза или слишком низкой точности классификации. Вместо одного сложного инструмента используются два более простых. В вышеупомянутом примере [17] применение теории нечетких множеств фактически добавляет к модели два элемента, которые можно правильно «настроить». Схема модели представлена на рис. 10.



Рис. 10. Три модуля нейронечеткой модели [17]

Первый модуль — это преобразование точных чисел в значения трех функций принадлежности. Второй — настроить сеть для прогнозирования этих трех значений с наименьшей ошибкой, а третий — повысить точность полученных прогнозов до точных цифр. Операции в каждом из этих трех модулей могут выполняться по-разному, поэтому для моделирования зависимой переменной можно использовать три инструмента, а не только саму искусственную нейронную сеть. Есть много примеров, когда гибридные модели, то есть те, в которых более одного инструмента используются совместно, дают более точные прогнозы, чем модели с одним инструментом [34–35]. Поэтому при анализе результатов исследований стоит рассмотреть возможность использования инструментов машинного обучения вместе с другими математическими инструментами.

Извлечение проверочного подмножества. Для обучения искусственных нейронных сетей из существующей базы данных выделяются три подмножества, содержащие как независимые, так и зависимые переменные: обучение, тестирование и проверка. Обучающее подмножество используется для обучения инструмента. Этот процесс продолжается до тех пор, пока MSE не перестанет уменьшаться в тестовой выборке, затем процесс обучения сети останавливается. Дальнейшее обучение сети могло бы привести к лучшему согласованию инструмента с обучающими данными, но возможности обобщения были бы потеряны (ошибки MSE для тестовой и проверочной выборки были бы намного выше) [30]. Такой эффект, называемый переобучением искусственной нейронной сети, схематично показан на рис. 12.

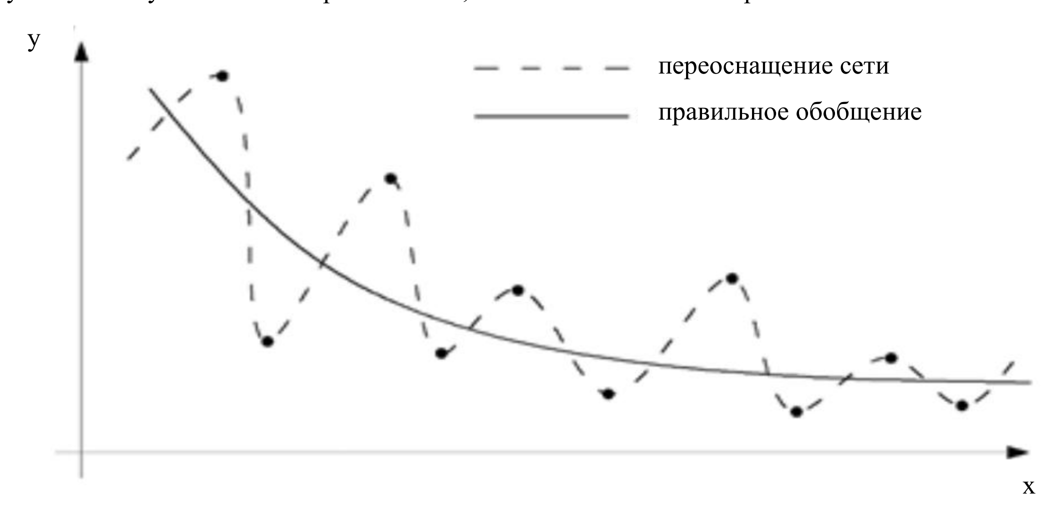


Рис. 12. Схематическое изображение переобучения искусственной нейронной сети и правильно обнаруженного тренда [30, 31]

Подмножество валидации используется для оценки качества прогнозов или классификации. Эти наборы независимых переменных и зависимая переменная не видят сеть в процессе обучения. Поэтому прогнозы (или классификации) делаются для проверочного подмножества, и путем сравнения с известными зависимыми переменными могут быть вычислены ошибки, описанные в предыдущих разделах. Очень важно указать, к какому подмножеству относится вычисленная ошибка. Один из методов оценки того, не переоборудован ли инструмент, — это сравнение ошибок (чаще всего MSE) для вышеупомянутых трех подмножеств, они должны быть на аналогичном уровне. Явно более низкая MSE для обучающего подмножества может указывать на переобучение, а явно более высокая MSE — на несовершенство инструмента. Явно более низкая MSE для тестовой и проверочной выборки предполагает необходимость выбора других параметров искусственной нейронной сети или других зависимых переменных (или даже другого инструмента).

Для больших баз данных подходит случайное разделение данных на три подмножества. В литературе встречаются предложения о том, что пропорции размера этих подмножеств должны быть в диапазоне 60:20:20–70:15:15 (преподаватель: тест: проверка) [36–37]. Для небольших баз данных случайный выбор подходит только в тех случаях, когда различные диапазоны зависимой переменной (или класса для классификации) одинаково численно представлены. Чаще всего это условие не выполняется. Тогда было бы хорошо обеспечить такое сбалансированное представление во всех трех подмножествах. Такая контролируемая разбивка данных использовалась в [11, 25]. В [11] из 66 испытаний стали одна марка (5 испытаний) была наименее многочисленной, а марка с 12 испытаниями — самой многочисленной. При случайном выборе тестов для подмножеств могло случиться так, что в обучающую подгруппу не было включено ни одного теста какой-либо марки стали, что, несомненно, помешало бы его автоматическому распознаванию. В [25] из 249 проанализированных дел только девять, по мнению авторов, следует отнести к категории «весьма вероятный сговор». После раздела набора данных на подмножества четыре из этих процедур были назначены обучающими, одна — тестовой и четыре — проверочными. Благодаря этой процедуре три из четырех процедур из набора для валидации могут быть правильно классифицированы.

В польской и англоязычной литературе можно найти примеры того, что авторы используют только концепции обучающего набора данных, тестового набора данных. Вероятно, это происходит по двум причинам. Некоторые компьютерные программы полностью (без вмешательства пользователя) контролируют, не переоборудован ли инструмент. Затем извлекается только проверочный набор (либо путем указания его количества или процентной доли в общих данных, либо путем выбора записей, которые будут использоваться для проверки). В этом случае подмножество проверки часто называют подмножеством тестов. Вторая причина

заключается в том, что некоторые инструменты машинного обучения (например деревья классификации, деревья C&RT[6]) защищены от переобучения другим способом, нежели контроль MSE для тестовой выборки. Качество этих инструментов можно проверить с помощью тестового подмножества (точно так же, как подмножество валидации — не участвуя в процессе обучения инструмента). Обычно уже при чтении работ по машинному обучению становится ясно, были выделены два или три подмножества данных, поэтому различная номенклатура набора, используемого для оценки качества работы, не является проблемой.

Проверка качества модели. Базы данных с небольшим количеством записей, используемые для построения модели на основе машинного обучения, делают модель менее устойчивой (т. е. дают существенно разные ошибки) при замене записей между подмножествами (обучение, тестирование и проверка). Чтобы узнать, работает ли построенная модель хорошо только с определенным разделением данных на подмножества, ее следует запустить на рандомизированных подмножествах. Используя предложения по пропорции разделения данных, их можно поделить на 5–7 подмножеств, и процессов построения модели должно быть выполнено как можно больше, так что на каждой итерации будет происходить обучение. Такая проверка качества модели называется перекрестной проверкой (на английском языке наборы данных при перекрестной проверке называются «свертками»). Ошибки MSE или MAPE затем усредняются. Если, однако, в каком-либо наборе данных ошибки значительно отклоняются от среднего значения, следует поискать причину этого.

Независимо от размера базы данных оценку ошибок также следует рассматривать через призму полезности результатов. Прогнозы с относительно большими ошибками, явления, которые нельзя точно описать, могут быть очень полезными и считаться важными. С другой стороны, прогнозы с теми же MSE или MAPE другого явления могут не дать какой-либо новой информации об изучаемом явлении. Поэтому, помимо числовых значений точности прогноза или точности классификации, важно ссылаться на само изучаемое явление.

Также важно проверить, не была ли построена другая модель для ранее изученного явления. Если да, то ссылка на эти предыдущие исследования (относительно уровня полученных там ошибок) также подтвердит качество нового патентованного решения. Если такие модели ранее не создавались, можно проверить качество нового решения, сравнив результаты с гораздо более простой моделью (например, на основе Microsoft Excel с надстройкой Solver).

Существует вопрос, на который пока нет четкого ответа: означает ли небольшое количество наборов данных, что полученные прогнозы, классификации и правила ассоциации неактуальны (именно из-за их небольшого количества)? Например, сложное и строгое правило со следующими индексами $sup = 0{,}01$, $conf = 1$, $lift = 100$ для базы данных с 10000 записями означает, что каждый раз, когда конкретный предшественник встречается 100 раз, указанный предшественник всегда (sic!) является последующим. То же правило для базы данных из 100 записей означает, что был только один уникальный случай, когда указанный предшественник имел место. Затем его сопровождал конкретный преемник. В одном случае нельзя говорить о правиле, но можно о случае (случайности). Однако, даже для небольшой базы данных (например, со 100 записями) правило 100-процентной достоверности, поддерживаемое тремя случаями, может уже указывать на повторяемость (при условии, что приращение для этого правила больше единицы).

Существуют небольшие базы данных обычно потому, что они малы сами по себе, потому что подготовка более крупных дорогостоящая, требует очень много времени, они к тому же могут быть следствием природы явления (например, ограниченное количество объектов, построенных одной и той же компанией). В реальности эти базы данных не могут быть расширены (в несколько раз или в несколько десятков раз). Изучение таких баз данных при соблюдении условий, описанных в этой главе, может привести к появлению новых, пока не обнаруженных зависимостей. Построенные, правильно функционирующие модели могут быть неприменимы напрямую к другим подобным явлениям, но они могут эффективно указывать на методы поиска общих взаимосвязей в случаях, когда количество наборов данных значительно больше.

Уникальные правила ассоциации или неожиданные автоматические классификации также могут указывать на области, на которых следует сосредоточить дальнейшие исследования описываемых явлений.

Заключение. Обсуждаемые в статье проблемы и вопросы, связанные с поиском взаимосвязей между многомерными входными и выходными данными, представлены в небольшом количестве случаев. Данная работа поэтому не может являться исчерпывающим обзором по теме. Объем статьи и ограниченный опыт автора в применении методов машинного обучения и интеллектуального анализа данных не позволяют описать большинство используемых методов. Проблемы при расчетах (в основном на небольших базах данных), содержащиеся в упомянутых в тексте работах, и их решения систематизированы таким образом, чтобы

[6] StatSoft. Internetowy Podręcznik Statystyki. URL: https://www.statsoft.pl/textbook/stathome.html (accessed 20.07.2020).

соблюдалась последовательность действий: от подготовки базы данных до расчетов, до обсуждения результатов. Невозможно четко указать, сколько записей в базе данных может свидетельствовать о «малой» или «большой» базе данных. Однако можно сказать, что для эффективного использования инструментов машинного обучения или интеллектуального анализа данных требуется как минимум несколько десятков записей. Применяя соответствующие процедуры (разработка входных данных, построение моделей), можно применять эти инструменты для успешного моделирования и изучения явлений, описанных всего в нескольких десятках случаев. На основе проведенных расчетов возможно сделать надежный вывод.

Библиографический список

1. Lissowski, G. Podstawy statystyki dla socjologów. Opis statystyczny. Tom 1 / G. Lissowski, J. Haman, M. Jasiński. — Warszawa: Wydawnictwo Naukowe Scholar, 2011. — 223 p.

2. Stanisławek J. Podstawy statystyki: opis statystyczny, korelacja i regresja, rozkłady zmiennej losowej, wnioskowanie statystyczne / J. Stanisławek. — Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2010. — 212 p.

3. Larose, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. 2nd ed. / D. T. Larose, C.D. Larose. — Hoboken, NJ, USA: Wiley-IEEE Press, 2016. — 309 p.

4. Larose, D. T. Metody I modele eksploracji danych / D.T. Larose. Warszaw: PWN, 2012. — 337 p.

5. Hand, D. Principles of Data Mining / D. Hand, H. Mannila, P. Smyth. — Cambridge, MA, USA: MIT Press, 2001. — 322 p.

6. Morzy, T. Eksploracja danych. Metody i algorytmy / T. Morzy. — Warszawa: PWN, 2013. — 533 p.

7. Bartkiewicz, W. Sztuczne sieci neuronowe. W: Zieliński JS. (red), Inteligentne systemy w zarządzaniu. Teoria i praktyka / W. Bartkiewicz. — Warszawa: PWN, 2000. — 348 p.

8. Rutkowski, L. Metody i techniki sztucznej inteligencji / L. Rutkowski. — Warszawa: PWN, 2012. — 449 p.

9. Doroshenko, A. Applying Artificial Neural Networks In Construction / A. Doroshenko // In: Proceedings of 2nd International Symposium on ARFEE 2019. — 2020. — Vol. 143. — P. 01029. https://doi.org/10.1051/e3sconf/202014301029

10. Feature Importance of Stabilised Rammed Earth Components Affecting the Compressive Strength Calculated with Explainable Artificial Intelligence Tools / H. Anysz, Ł. Brzozowski, W. Kretowicz, P. Narloch // Materials. — 2020. — Vol. 13. — P. 2317. https://doi.org/10.3390/ma13102317

11. Artificial Neural Networks in Classification of Steel Grades Based on Non-Destructive Tests / A. Beskopylny, A. Lyapin, H. Anysz, et al. // Materials. — 2020. — Vol. 13. — P. 2445. https://doi.org/10.3390/ma13112445

12. Anysz, H. Wykorzystanie sztucznych sieci neuronowych do oceny możliwości wystąpienia opóźnień w realizacji kontraktów budowlanych / H. Anysz. — Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 2017. — 280 p.

13. Rabiej, M. Statystyka z programem Statistica / M. Rabiej. — Poland: Helion, Gliwice, 2012. — 344 p.

14. Mrówczyńska, M. Compression of results of geodetic displacement measurements using the PCA method and neural networks / M. Mrówczyńska, J. Sztubecki, A. Greinert // Measurement. — 2020. — Vol. 158. — P. 107693. https://doi.org/10.1016/j.measurement.2020.107693

15. Mohamad-Saleh, J. Improved Neural Network Performance Using Principal Component Analysis on Matlab / J. Mohamad-Saleh, B. C. Hoyle // International Journal of the Computer, the Internet and Management. — 2008. — Vol. 16. — P. 1–8.

16. Juszczyk, M. Application of PCA-based data compression in the ANN-supported conceptual cost estimation of residential buildings / M. Juszczyk // AIP Conference Proceedings. — 2016. — Vol. 1738. — P. 200007. https://doi.org/10.1063/1.4951979

17. Anysz, H. Neuro-fuzzy predictions of construction site completion dates / H. Anysz, N. Ibadov // Technical Transactions. Civil Engineering. — 2017. — Vol. 6. — P. 51–58. https://doi.org/10.4467/2353737XCT.17.086.6562

18. Rogalska, M. Wieloczynnikowe modele w prognozowaniu czasu procesów budowlanych / M. Rogalska. — Lublin: Politechniki Lubelskiej, 2016. — 154 p.

19. Kaftanowicz, M. Multiple-criteria analysis of plasterboard systems / M. Kaftanowicz, M. Krzemiński // Procedia Engineering. — 2015. — Vol. 111. — P. 351–355. https://doi.org/10.1016/j.proeng.2015.07.102

20. Anysz, H. The influence of input data standardization method on prediction accuracy of artificial neural networks / H. Anysz, A. Zbiciak, I. Ibadov // Procedia Engineering. — 2016. — Vol. 153. — P. 66–70. https://doi.org/10.1016/j.proeng.2016.08.081

21. Nicał, A. The quality management in precast concrete production and delivery processes supported by association analysis / A. Nicał, H. Anysz // International Journal of Environmental Science and Technology. — 2020. — Vol. 17. — P. 577–590. https://doi.org/10.1007/s13762-019-02597-9

22. Anysz, H. The association analysis for risk evaluation of significant delay occurrence in the completion date of construction project / H. Anysz, B. Buczkowski // International Journal of Environmental Science and Technology. — 2019. — Vol. 16. — P. 5396–5374. https://doi.org/10.1007/s13762-018-1892-7

23. Zeliaś, A. Prognozowanie ekonomiczne. Teoria, przykłady, zadania / A. Zeliaś, B. Pawełek, S. Wanat. — Warszawa: PWN, 2013. — 380 p.

24. Juszczyk, M. Modelling Construction Site Cost Index Based on Neural Network Ensembles/ M. Juszczyk, A. Leśniak // Symmetry. — 2019. — Vol. 11. — P. 411. https://doi.org/10.3390/sym11030411

25. Anysz, H. Comparison of ANN Classifier to the Neuro-Fuzzy System for Collusion Detection in the Tender Procedures of Road Construction Sector / H. Anysz, A. Foremny, J. Kulejewski // IOP Conference Series: Materials Science and Engineering. — 2019. — Vol. 471. — P. 112064. https://doi.org/10.1088/1757-899X/471/11/112064

26. Piegorsch, W. W. Confusion Matrix. In: Wiley StatsRef: Statistics Reference Online. — 2020. — P. 1–4. https://doi.org/10.1002/9781118445112.stat08244

27. Kot, S. M. Statystyka / S. M. Kot, J. Jakubowski, A. Sokołowski. — Warszawa: DIFIN, 2011. — 528 p.

28. Aczel, A. D. Statystyka w zarządzaniu / A. D. Aczel, J. Saunderpandian. — Warszawa: PWN, 2000. — 977 p.

29. Narloch, P. Predicting Compressive Strength of Cement-Stabilized Rammed Earth Based on SEM Images Using Computer Vision and Deep Learning / P. Narloch, A. Hassanat, A. S. Trawneh, et al. // Applied Sciences, 2019. — Vol. 9. — P. 5131. https://doi.org/10.3390/app9235131

30. Tadeusiewicz, R. Sieci neuronowe / R. Tadeusiewicz. — Kraków: Akademicka Oficyna Wydawnicza, 1993. — 130 p.

31. Anysz, H. Designing the Composition of Cement Stabilized Rammed Earth Using Artificial Neural Networks / H. Anysz, P. Narloch // Materials. — 2019. — Vol. 12. — P. 1396. https://doi.org/10.3390/ma12091396

32. Zadeh, L. A. Fuzzy Sets / L. A. Zadeh // Information and Control. — 1965. — Vol. 8. — P. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

33. Yagang Zhang. A hybrid prediction model for forecasting wind energy resources / Yagang Zhang, Guifang Pan // Environmental Science and Pollution Research. — 2020. — Vol. 27. — P. 19428–19446. https://doi.org/10.1007/s11356-020-08452-6

34. Eugene, E.A. Learning and Optimization with Bayesian Hybrid Models. 2020 American Control Conference (ACC) / E. A. Eugene, Xian Gao, A. W. Dowling. — IEEE. — 2020. https://doi.org/10.23919/ACC45564.2020.9148007

35. Neural Network Design / M. T. Hagan, H. B. Demuth, M. H. Beale, O. De Jesús. — Martin Hagan: Lexington, KY, USA, 2014. — 1012 p.

36. Osowski, S. Sieci neuronowe do przetwarzania informacji / S. Osowski. —Warszawa: Oficyna Wydawnicza PW, 2006. — 419 p.

Поступила в редакцию 01.10.2021
Поступила после рецензирования 25.10.2021
Принята к публикации 26.10.2021

Об авторе:

Аныш, Хуберт, старший преподаватель факультета гражданского строительства Варшавского технологического университета (Польская Республика, 00-661, г. Варшава, пл. Политехники, 1), доктор философии, Scopus, Researcher, ORCID, h.anysz@il.pw.edu.pl

Автор прочитал и одобрил окончательный вариант рукописи.